DOCUMENT RESUME

ED 437 410 TM 030 562

AUTHOR van der Linden, Wim J.; Carlson, James E.

TITLE Calculating Balanced Incomplete Block Design for Educational Assessments.

INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.

REPORT NO RR-99-08
PUB DATE 1999-00-00
NOTE 25p.; Portions of the paper presented at the National Assessment Governing Board Achievement Levels Workshop (Boulder, CO, August 20-22, 1997).
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; *Item Banks; *Linear Programming; *Research Design; Research Methodology
IDENTIFIERS *Balanced Incomplete Block Spiralling; Large Scale Assessment; National Assessment of Educational Progress

ABSTRACT

A popular design in large-scale educational assessments is the balanced incomplete block design. The design assumes that the item pool is split into a set of blocks of items that are assigned to assessment booklets. This paper shows how the technique of 0-1 linear programming can be used to calculate a balanced incomplete block design. Several structural as well as practical constraints on this type of design are formulated as linear (in)equalities. In addition, possible objective functions to optimize the design are discussed. The technique is demonstrated using an item pool from the 1996 Grade 8 Mathematics National Assessment of Educational Progress Project. (Contains 2 tables and 16 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 437 410

TM

Calculating Balanced Incomplete Block Design for Educational Assessments

Research Report 99-08

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Wim J. van der Linden, University of Twente
James E. Carlson, CTB/McGraw-Hill

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030562

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of Educational Measurement and Data Analysis

BEST COPY AVAILABLE

Calculating Balanced Incomplete Block Design for Educational Assessments

Wim J. van der Linden
University of Twente

James E. Carlson
CTB / McGraw-Hill

Abstract

A popular design in large-scale educational assessments is the balanced incomplete block design. The design assumes that the item pool is split into a set of blocks of items that are assigned to assessment booklets. This paper shows how the technique of 0-1 linear programming can be used to calculate a balanced incomplete block design. Several structural as well as practical constraints on this type of design are formulated as linear (in)equalities. In addition, a variety of possible objective functions to optimize the design are discussed. The technique is demonstrated using an item pool from the 1996 Grade 8 Mathematics NAEP Project.

Calculating Balanced Incomplete Block Designs for Educational Assessments

The purpose of an educational assessment is to evaluate the performances of a population of students on a pool of test items representative of a given subject area. Typically, the population and the pool are too large to involve all students in the assessment or to give all items to each student. An obvious strategy, therefore, is to sample students and items.

Typically, sampling of students takes place through a complex probabilistic, multistage sampling plan involving several levels of units. A description of the sampling plan used for sampling students in the National Assessment of Educational Progress (NAEP) is given in Rust and Johnson (1992).

When educational assessments were still based on classical test theory, items were also sampled randomly. The parameter of interest were the mean scores of the population of students on the individual items in the pool. An efficient strategy for estimating these parameters is multiple-matrix sampling. In multiple-matrix sampling, both the students and the items are sampled randomly assigning subsets of items to subsets of students (Sirotnik, 1974). An important result on multiple-matrix sampling was given in Lord (1962; see also Lord & Novick, 1968, sect. 11.12) who showed that the mean scores of a population of students on a pool of items are estimated best if each single item is administered to a random, nonoverlapping subset of students. In practice, this design is not feasible because of the complicated logistics involved in delivering single items to examinees, but it served as an important benchmark when classical sampling procedures for educational assessments were designed.

With the advent of item response theory (IRT), the interest in educational assessments shifted from mean scores on individual items to the full population distribution on the ability parameter in the model. One of the features of IRT helpful in educational assessments is that, though different item-student combinations yield different

statistical precision, random assignment of items to students is no necessary condition for consistent estimation of the ability distribution. Hence, a feasible approach is to assemble assessment booklets from an item pool according to some practical principle and assign them to students in units sampled at the lowest level of the population.

Both in the National Assessment of Educational Progress (NAEP) in the USA and in the Dutch Periodiek Peilingsonderzoek van het Onderwijs (PPON) projects, tests are assembled following the structure of a balanced incomplete block (BIB) design (Johnson, 1992; Wijnstra, 1988). The design assumes that the pool of items is split into a set of blocks. The split need not be random but may be based on such practical issues as the wish to offer students blocks with stimulating combinations of items or to match blocks across booklets with respect to the time needed to complete them. Also, the number of booklets that have to be designed is predetermined. Finally, booklets are spiraled across students in the lowest unit (usually school classes) to minimize the cluster effects involved in sampling a hierarchically structured population.

In a BIB design, the assignment of blocks to assessment booklets is controlled by the following constraints:

1. The number of blocks assigned to each booklet is between certain bounds.
2. The number of booklets each block is assigned to is between certain bounds.
3. Combinations of blocks are assigned to a minimum number of booklets.

This set of constraints will be referred to as structural constraints. The third type of constraint is needed only if statistical relations between items in different blocks, for example, their covariances, have to be estimated. Figure 1 gives an example of a BIB design which is derived from Johnson (1992, Fig. 1).

[Figure 1 about here]

If no other constraints had to be imposed on BIB designs, the actual assignment of the blocks to the assessment booklets would be a simple task. As the example in Figure 1 suggests, a procedure in which the blocks are systematically rotated across the booklets would already do. However, in practice several additional constraints, for example, on item content, format, and response time, may have to be imposed on the composition of the booklets. Such constraints will be referred to as practical constraints. If both structural and practical constraints are to be imposed on the assignment of the blocks to the booklets, the assignment process quickly becomes too complicated for manual execution. The same conclusion holds if the assignment has to be optimized with respect to some objective, for instance, an important psychometric aspect.

The purpose of this paper is to show how the technique of 0-1 linear programming (LP) can be exploited to assemble optimal sets of booklets following a BIB design. In the remainder of this paper, first several practical constraints on BIB design and possible objective functions are discussed. Then, a general 0-1 LP model for assembling booklets from a pool of blocks is introduced. The paper concludes with an empirical example in which a pool of blocks from 1996 Grade 8 Mathematics NAEP Project was used to assemble an optimal set of assessment booklets.

Some Practical Constraints and Objective Functions

Practical constraints on test assembly can be classified in various ways. A convenient classification it is the following (van der Linden, 1998):

1. Constraints based on categorical item attributes, such as item content, format, cognitive level, and whether or not an item has graphics. Each categorical attribute partitions the item pool, and constraints on these attributes specify a desired distributions of items over the partition.
2. Constraints based on quantitative item attributes, that is, on parameters or coefficients with numerical values, such as item p-values, word counts, and

(expected) response times. Quantitative constraints require sums or averages of attributes values to be between certain bounds.

3. Logical (or Boolean) constraints to deal with certain dependencies between the items in the pool. Two important cases are items organized around common stimuli ("item sets") and items that can not be in the same form because of content overlap ("enemies").
4. Constraints to set the length of the test form or some of its sections to a prespecified number of items.

Examples of each of these types of constraints are given in the general 0-1 LP model for calculating BIB designs below.

If assessment booklets are assembled from a set of blocks, the main focus may be on the constraints in the first two categories. The constraints in the third category are relevant, for example, if items in different blocks are enemies. If so, the blocks should be treated as enemies themselves. Item sets only occur within blocks and therefore need no special concern when blocks are combined into booklets. Finally, if the blocks are matched on the time needed to complete them, the constraints on test length in the last category boil down to those on the number of blocks per booklet. An alternative to matching blocks on time is to leave the number of items per block free, use these numbers as an attribute, and constrain their sum per booklet.

Possible Objective Functions

The technique of 0-1 LP can be used to find a design satisfying a full set of constraints. In mathematical programming, solutions that meet the full set of constraints are known as feasible solutions. An objective function is used to identify an optimum in the set of feasible solutions. If the goal is only to find a BIB design and there exist no further preferences, all feasible solutions are equally good. In this case, an arbitrary objective function defined on (a subset of) the decision variables will do. However, the

objective function can also be used to optimize the design with respect to an important psychometric aspect.

The following possible objective are suggested:

1. Minimization of a suitable function of the covariance matrix of the (MML) estimators of the parameters characterizing the population distributions, such as their determinant or trace. This objective makes sense if multiple distributions have to be evaluated and booklets have to be optimized with respect to different distributions (see below).
2. If the interest is not only in estimating properties of the distributions of certain populations but also in reporting individual scores to schools, it may be helpful to increase the efficiency of the individual ability estimators maximizing the booklet information functions over well-chosen intervals. A favorable side effect of this objective function is that the improved estimation of the individual θs increases the robustness of marginal analyses of group differences against model misspecifications (Mislevy, Beaton, Kaplan & Sheehan, 1992).
3. Student motivation to answer the items in the assessment can be expected to be low if their probabilities of success on the items are consistently low or high. An objective function can be chosen that minimizes the distances between target values and the actual probabilities on the items for ability values typical of the subpopulations of students the booklets are administered to.
4. If assessment tests are speeded, too many items may not be reached. If estimates of the time needed to complete the items are available for the various clusters of students, it may make sense to use an objective function that optimizes the match between the items and the students they are administered too.

For the mainstream IRT models, the above functions of the covariance matrix in the first objective are nonlinear in the items. Therefore, application of the technique of 0-1 LP requires that a good linear approximation be available. This strategy has been possible in another multi-parameter IRT test assembly problem (van der Linden, 1996) but has not yet been explored for the current problem. The second objective has been used in a variety of other test assembly problems (van der Linden, 1998); its application to the problem of assembling assessment booklets does not involve any new aspects. The third objective function will be used in the empirical example below. The fourth objective function is possible if the items have been pretested to obtain empirical estimates of their response time distributions or if good subjective estimates exist.

To implement the objectives, prior knowledge about the students is needed. The last three objectives seek an optimal match between the attributes of the items and characteristics of the students. If these characteristics are not directly known, they can be predicted from background variables, which are also needed to define relevant strata and clusters in the sampling plan, provided the necessary regression functions are known, for example, from a previous assessment.

As already noted, the first objective makes sense if the distributions of multiple subpopulations have to be evaluated. These subpopulations are generally defined using background variables. Empirical priors for the parameters of their distribution functions may be derived from previous assessments. The idea is to assemble the booklets while optimizing the efficiency of the covariance matrix with respect to the priors for the distribution parameters.

Background variables can also be used to match units in the sample. It is assumed throughout this paper that the booklets are administered to subgroups of units matched on relevant background variables. In addition, since the assembly of each of the booklets may have to be optimized with respect to these subpopulations, special objective functions are needed to guarantee a solution that is simultaneously optimal for all subpopulations. In the

example in this paper, an objective function based on the maximin criterion is used for this purpose.

0-1 LP Model for Balanced Incomplete Block Designs

A general framework for a 0-1 LP model for balanced incomplete block designs is presented. It is assumed that the items have been calibrated previously using the 3-parameter logistic (3PL) model:

$$P_i(+|\theta) = c_i + (1-c_i)\{1+\exp[-a_i(\theta-b_i)]\}^{-1}, \quad (1)$$

where $a_i \in (0,\infty)$, $b_i \in (-\infty,\infty)$, and $c_i \in [0,1]$ are the discrimination, difficulty and, guessing parameter for item i, respectively (e.g., Lord, 1980). In addition, the following notation is needed.

The individual blocks in the pool are represented by indices j=1,...,N. To represent pairs of blocks a second index k with the same range of possible values is used. Booklets are denoted by b=1,...,B. Binary variables x_{jb} are used to decide whether (x_{jb}=1) or not (x_{jb}=0) block j is assigned to booklet b. Likewise, binary variables z_{jkb} are used to assign pair (j,k) to booklet b. Special constraints will be formulated below to keep the values of these two categories of variables consistent.

The distribution of blocks across booklets is described by the following numbers:

c_1: number of blocks per booklet;

c_2: number of booklets per block;

c_3: minimum number of booklets per pair of blocks.

To illustrate the possibility to control the contents of the booklets beyond these numbers, three different kinds of additional constraints are introduced. First, it is assumed that the blocks are classified by content. Content is represented by a categorical attribute

c=1,...,C, where V_c is defined as the subset of blocks in the pool belonging to content category c and n_c is the number of blocks to be selected from V_c. Second, to illustrate the treatment of a categorical attribute it is assumed that the booklets have to be controlled for response time. The response time permitted for block j is denoted as q_j, whereas the total amount of time permitted for booklet b is T_b. Finally, it is assumed that some blocks are "enemies" in the sense that they can not be assigned to the same booklet. The sets of indices of enemies are denoted by V_e, e=1,...,E.

As an example of an objective function, the case of minimization of the distances between the probabilities of success on the items and their target values is used. Let τ_b be the target for the success probabilities on the items in booklet b, and θ_b^* a typical ability value for the students for which booklet b is designed. Finally, the set of indices of the items in block j is denoted as V_j and it is assumed that block j has n_j items.

The model is as follows:

minimize y (objective function) **(2)**

subject to

$$[n_j^{-1} \sum_{i \in V_j} P_i(+|\theta_b^*) - \tau_b] x_{jb} \leq y, \quad b=1,...,B,\ j=1,...,N, \quad \text{(success probabilities)} \quad \textbf{(3)}$$

$$[n_j^{-1} \sum_{i \in V_j} P_i(+|\theta_b^*) - \tau_b] x_{jb} \geq -y, \quad b=1,...,B,\ j=1,...,N, \quad \text{(success probabilities)} \quad \textbf{(4)}$$

$$\sum_{j=1}^{N} x_{jb} = c_1, \quad b=1,...,B, \quad \text{(\# blocks per booklet)} \quad \textbf{(5)}$$

$$\sum_{b=1}^{B} x_{jb} \le c_2, \quad j=1,...,N, \qquad \text{(\# booklets per block)} \quad \textbf{(6)}$$

$$\sum_{b=1}^{B} z_{jkb} \ge c_3, \quad j<k=1,...,N, \qquad \text{(\# booklets per pair)} \quad \textbf{(7)}$$

$$x_{jb} + x_{kb} \ge 2z_{jkb}, \quad j<k=1,...,N,\ b=1,...,B, \qquad \text{(consistent assignment)} \quad \textbf{(8)}$$

$$\sum_{b=1}^{B} \sum_{j \in V_c} x_{jb} \ge n_c, \quad c=1,...,C, \qquad \text{(content)} \quad \textbf{(9)}$$

$$\sum_{j=1}^{N} q_j x_{jb} \le T_b, \quad b=1,...,B, \qquad \text{(response time)} \quad \textbf{(10)}$$

$$\sum\sum_{(j<k) \in V_e} z_{jkb} \le 1, \quad e=1,...,E,\ b=1,...,B, \qquad \text{(enemies)} \quad \textbf{(11)}$$

$$x_{jb} \in \{0,1\}, \quad j=1,...,N,\ b=1,...,B, \qquad \text{(definition of } x_{jb}\text{)} \quad \textbf{(12)}$$

$$z_{jkb} \in \{0,1\}, \quad j<k=1,...,N,\ b=1,...,B. \qquad \text{(definition of } z_{jkb}\text{)} \quad \textbf{(13)}$$

The constraints in (3)-(4) require the sum of the differences between the targets and the actual success probabilities to be in the interval [-y,y]. The size of this interval is minimized in the objective function in (1). The constraints in (5)-(6) define the size of the booklet in terms of the numbers of blocks and the number of times a block is assigned to

a booklet, respectively, whereas (7) sets the minimum number of booklets to which each possible pair is assigned equal to c_3. The constraints in (8) stipulate that each time a pair of blocks is assigned ($z_{jkb}=1$), it also holds that the individual blocks are assigned ($x_{jb}=1$ and $x_{kb}=1$). Observe that the reverse implication is not necessary. However, if the reverse implication is desired, the following constraints should be added to the model:

$$z_{jb} + z_{kb} - 1 < z_{jkb}, \quad j<k=1,...,N,\ b=1,...,B. \qquad \text{(consistent assignment)} \quad \textbf{(14)}$$

Due to the constraints in (9), at least n_c blocks from content category are assigned to a booklet, while the constraints in (10) guarantee that for booklet b no more than T_b minutes are needed. The constraints in (11) prevent from assigning more than one block from each set of enemies. Finally, the constraints in (12)-(13) define the ranges of the decision variables

The objective function in (1), along with the constraints in (2)-(3), is of the maximin type. It minimizes the maximum deviation between the targets and success probabilities across all booklets. As indicated earlier, if the interest is only in calculating a feasible solution for the set of constraints in (4)-(13), this objective function can be replaced by any arbitrary linear function of the decision variables in the model, for example, their sum.

The number of variables in this problem is equal to BN[1+(N-1)/2]+1, namely BN variables x_{jb}, BN(N-1)/2 variables z_{jkb} and one variable y in the objective function. The number of constraints in the core of the model (Equations 3-8) is equal to (B+1)N(N-1)/2+B(2N+1)+N. In the empirical example below, B was equal to 26 and N to 13, yielding a model with 2,367 variables. For problems of this size, a heuristic for solving 0-1 LP problems is needed, for example, one of the heuristics available in ConTEST (Timminga, van der Linden, & Schweizer, 1996) or in CPLEX (ILOG, 1998).

Empirical Example

The goal of this example was to provide a post hoc illustration of the technique using a pool of item blocks from the 1996 NAEP Grade 8 Mathematics Project (Reese, Miller, Mazzeo & Dossey, 1997). The pool consisted of 13 blocks of dichotomous and polytomous items which had been combined in 26 booklets in the NAEP assessment. All dichotomous items were calibrated using the 3PL model in (1) for the dichotomous items and the generalized partial credit model for the polytomous items (Muraki, 1992). In all, the pool had 139 dichotomous and 25 polytomous items. The following five scales were needed to calibrate the item pool: (1) Number, Sense, and Operations; (2) Measurement; (3) Geometry and Spatial Sense; (3), Data Analysis, Statistics, and Probability; and (5) Algebra and Functions.

The model used to calculate an optimal balanced incomplete block design was the one in (2) through (8) with the definitions of the decision variables in (12) and (13). An objective function was formulated to select the blocks to have items with probabilities of success as closely as possible to .50 on the dichotomous items for typical ability values in the subpopulations of students. For the polytomous items, the differences between the expected scores and the midpoint of their score intervals were minimized. To remove the effects of scale differences between the polytomous and dichotomous scores in (3) and (4), the expected scores and midpoints on the polytomous items were first scaled back to [0,1]). The subpopulations were fictitious; they were chosen to be functioning at the 25th, 50th and 75th percentile of the national distributions on the five mathematics scales in the 1996 NAEP assessment.

More specifically, the model was as follows:

1. In the constraints in (3) and (4), the ability values for the target populations, θ^*, and the values for the target probabilities (for the polytomous items: target expected scores), τ, were substituted.
2. The total number of booklets assembled was equal to 26. Ten booklets were

assembled for the target population at the 50th percentile and eight booklets for each of the populations at the 25th and 75th percentile.

3. In the constraints in (5), the number of blocks per booklet was set equal to three.
4. In the constraints in (6), an upper limit of six booklets was imposed on the number of times a block could be assigned to a different booklet.
5. In the constraints in (7), the number of times each pair of blocks was assigned to a common booklet was set equal to at least once.

The specifications for the numbers of blocks and booklets were the regular specifications used in the 1996 assessment. Similarly, like the 1996 assessment, no further constraints on booklet content or any block or item attributes were imposed. The total number of decision variables and constraints in the model were equal to 2,367 and 2,197.

The model was solved using the CPLEX software (ILOG, 1998) on a PC with Pentium Pro/166MHz processor. As already noted, problems of this size are large for the search algorithms implemented in CPLEX. The approach was therefore to stop the algorithm when it did no longer succeed in finding (integer) solutions with improved values for the objective function (in this example after 55 hours). The best solution obtained at this point of time is given in Figure 2. The value for the objective function

[Figure 2 about here]

associated with the solution was .3211. That is, for none of the items the absolute difference between the actual probability of success (expected relative scores) and its target was larger than this value. Also, the mean absolute difference across items was calculated; it was equal to .2193.

Thus, though the subpopulations were chosen to have abilities varying as widely as between the 25th and 75th percentile in the national distributions on the five mathematics scales, a design was found for which none of the items had differences between the probabilities of success (expected relative scores) smaller than .1789 or larger than .8211

for any of these subpopulations. Across all items, the average difference for each subpopulation was not smaller than .2807 or larger than .7193.

Concluding Remark

An important assumption in this paper is that the item pool is already organized into blocks of items. Though this assumption is based on current practice, the existence of blocks by itself is a rather stringent constraint on the assembly of the assessment booklets. This point can easily be demonstrated for the objective function in the empirical example in this paper. If the items in the blocks happen to vary considerably in difficulty, it will never be possible to assign the blocks to subpopulations for which the objective function yields low values. But even if the blocks are homogeneous in difficulty, some difficulty levels may be over- or underrepresented and no favorable result is guaranteed.

In principle, it is possible to assign items directly to assessment booklets for subpopulations. The problem then boils down to an instance of multiple-form test assembly (van der Linden & Adema, 1998), with special constraints to guarantee a balanced-incomplete-block structure among the set of forms. These constraints are direct generalizations from those in (5)-(8).

The reason items in educational assessments are often pre-assembled into blocks is to neutralize possible differences in context effects of the items among students who receive different forms. On the other hand, assembly of assessment booklets directly from the items in the pool is likely to result in designs that are better in terms of the objective function used in the assembly process. Whether or not pre-assembly of item blocks should be recommended ultimately depends on the tradeoff between these two factors.

References

Johnson, E. G. (1992). The design of the national assessment of educational progress. Journal of Educational Measurement, 29, 95-110.

ILOG (1998). CPLEX 6.0 Documentation supplement [Computer software]. Incline Village, NV: ILOG, Inc.

Lord, F. M. (1962). Estimating norms by item sampling. Educational and Psychological Measurement, 22, 259-267.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 133-162.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. Applied Psychological Measurement, 17, 351-363.

Rust, K. F, & Johnson, E. G (1992). Sampling and weighting in the national assessment. Journal of Educational Measurement, 17, 111-129.

Sirotnik, K. (1974). An introduction to matrix sampling for the practitioner. In W. J. Popham (Ed.), Evaluation in education: Current applications (453-399). Berkeley, CA: McCutchen.

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1996). ConTEST [Computer program and manual]. Groningen, The Netherlands: iec ProGAMMA.

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). NAEP 1996 mathematics report card for the nation and the states. Washington, DC: National Center

for Education Statistics.

van der Linden, W. J. (1996). Assembling test for the measurement of multiple traits. Applied Psychological Measurement, 20, 373-388.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. Applied Psychological Measurement, 22, 195-211.

van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. Journal of Educational Measurement, 35, 185-198. [Addendum in Journal of Educational Measurement, 36, 90-91]

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.

Wijnstra, J. M. (Ed.) (1988). Balans van het rekenonderwijs in de basisschool: Uitkomsten van de eerste rekenpeiling medio en einde basis onderwijs [Assessment of artihmetic in elementary education: Results from the first study in elementary education]. Arnhem, The Netherlands: Cito.

Authors' Note

This paper was prepared while the first author was on a sabbatical leave at the Law School Admission Council, Newtown, Pennsylvania and the second author was at Educational Testing Service. Portions of the paper were presented at the National Assessment Governing Board Achievement Levels Workshop, Boulder, CO, August 20-22, 1997. The authors are indebted to Wim M. M. Tielen for his computational assistance.

Table 1. Example of a balanced incomplete block design (seven blocks; seven booklets; each possible pair of blocks in one booklet)

Booklet	Blocks
1	A B D
2	B C E
3	C D F
4	D E G
5	E F A
6	A G B
7	B A C

Table 2. Balanced incomplete block design calculated for the 1996 NAEP Grade 8 Mathematics Project (13 blocks; Booklet 1-8 for subpopulation at 25th, Booklet 9-18 for subpopulation at 50th, and Booklet 19-26 for subpopulation at 75th percentile)

Booklet	Blocks	Booklet	Blocks
1	D E I	14	E L M
2	D H K	15	C J L
3	B D L	16	C D J
4	A E H	17	B G I
5	D J M	18	C D F
6	G K L	19	B C G
7	A C I	20	A B M
8	A D G	21	C G M
9	B E J	22	B E J
10	G H J	23	I K M
11	F H M	24	H I L
12	A J K	25	A F L
13	E F G	26	F I J

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

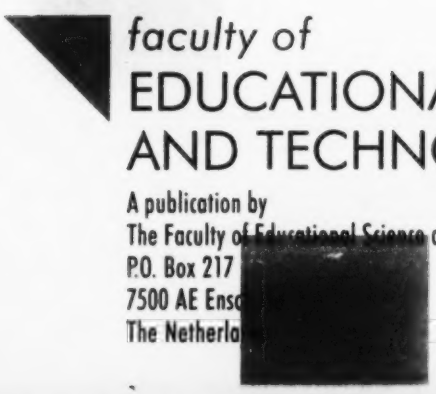
RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.



faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Ens
The Netherla

BEST COPY AVAILABLE